Maximum Likelihood Estimation of Higher-Order Integer-Valued Autoregressive Processes

Ruijun Bu, Brendan McCabe*
The University of Liverpool

Kaddour Hadri
Durham University

Final Version
(April 2008)

*Corresponding author: Management School, Chatham Street, Liverpool, L69 7ZH, UK, Tel: +44-151-795-3705, Fax: +44-151-795-3005, Email: Brendan.McCabe@liv.ac.uk (B. McCabe).

Maximum Likelihood Estimation of Higher-Order Integer-Valued Autoregressive Processes

Abstract

In this paper, we extend earlier work of Freeland and McCabe (2004) and develop a general framework for maximum likelihood (ML) analysis of higher-order integer-valued autoregressive processes. Our exposition includes the case where the innovation sequence has a Poisson distribution and the thinning is Binomial. A recursive representation of the transition probability of the model is proposed. Based on this transition probability, we derive expressions for the score function and the Fisher information matrix, which form the basis for maximum likelihood estimation and inference. Similar to the results in Freeland and McCabe (2004), we show that the score function and the Fisher information matrix can be neatly represented as conditional expectations. Using the $INAR(2)$ specification with Binomial thinning and Poisson innovations, we examine both the asymptotic efficiency and finite sample properties of the ML estimator in relation to the widely used conditional least squares (CLS) and Yule-Walker (YW) estimators. We conclude that, if the Poisson assumption can be justified, there are substantial gains to be had from using ML especially when the thinning parameters are large.

Keywords: Time series of counts; $INAR(p)$ model; Poisson autoregressive models; Maximum Likelihood Estimation; Asymptotic relative efficiency.

1. Introduction

Recently there has been growing interest in modeling time series of small counts that arise in various fields of statistics. Examples include the number of customers waiting to be served at a counter recorded at discrete points in time; the daily number of absent workers in a firm; the monthly cases of rare infectious diseases in a specified area; the monthly number of claimants collecting wage loss benefit for injuries in the workplace and so on. Typically, such time series take on only small non-negative integer values and often exhibit short-range dependence. Traditional continuous variable models are apparently inappropriate in that they would invariably produce non-integer forecast values. As a result, some specific class of time series models has to be entertained to explicitly account for the discreteness. This paper is concerned with a special class of observation-driven models called integer-valued autoregressive processes introduced independently by Al-Osh and Alzaid (1987) and McKenzie (1988). In this paper we use the notation $INAR(p)$ to mean that the thinning operator (with p lags) of the process is Binomial while the discrete innovations process is left unspecified. When the innovations process is specified to be Poisson then we write $INAR(p)$-P. If neither the thinning nor the innovations processes are fully specified we use the generalised $GINAR(p)$ notation.

Estimation of $INAR(p)$ process can be carried out in a variety of ways. Common ways for estimating parameters include the method of moments (based on the Yule-Walker (YW) equations) and conditional least squares (CLS). The implementation of both approaches is relatively simple and they are asymptotically equivalent. Al-Osh and Alzaid (1987) showed how maximum likelihood (ML) can be implemented for estimating the parameters of the $INAR(1)$-P model i.e. when Binomial thinning is used and the innovation sequence is assumed to be Poisson. They compared the finite sample properties of the three estimation methods and concluded that ML is worth the extra calculation because of the gain in terms of the bias and the mean squared error (MSE). Freeland and McCabe (2004) (FM) also considered the ML framework and derived new expressions for the score and information matrix as well as deriving a general test of specification for the model. However, both the work of Al-Osh and Alzaid (1987) and FM are confined to the first-order model. Recently, however, Jung and Tremayne (2006) considered estimation of the $INAR(2)$ model using the method of moments. Drost et al (2008) consider one-step asymptotically efficient estimation of the $INAR(p)$-P model citing computational difficulties with the convolutions involved in the ML

method. The main contribution of the present paper is to extend earlier work and develop a general framework for likelihood analysis of higher order $GINAR(p)$ processes with general thinning operators and innovation distributions. We derive the likelihood, using a recursive formulation of the transition probabilities, which facilitates both numerical computations and the derivative calculations required for the score and information quantities. Similar to the results in FM, we show that all elements of the score and the Fisher information matrix can be represented in terms of conditional expectations which enhances the interpretation of these quantities. While the results are quite general, we also specialise to the situation where the thinning processes are Binomial and the innovation sequence is Poisson and provide specific formulae for computational use in this case. The additional distributional assumptions also allow for verifiable conditions to ensure the existence of asymptotic stationary and limit distributions for the process and associated estimators. We also investigate the asymptotic relative efficiency (ARE) of CLS to ML and these calculations show that there are quite substantial efficiency gains to be had by imposing the Poisson assumption (should it be justified) especially when the thinning parameters are large in magnitude. A Monte Carlo study shows that ML also has advantages in small samples in terms of bias and mean squared error (MSE).

The remainder of the paper is organized as follows. Section 2 presents a likelihood framework for the $GINAR(p)$ model and derives the likelihood, score and information. In Section 3, we outline the $INAR(p)$-P process and briefly review its main statistical properties. In Section 4, the asymptotic relative efficiency of the ML estimator is examined and a simulation experiment looks at the small sample properties. Section 5 concludes. The proofs and other details are contained in Appendices.

2. Likelihood Calculations for the $GINAR(p)$ Model

In this section we consider a $GINAR(p)$ model where the thinning operators and arrivals distribution are specified in just enough detail to enable the likelihood and the associated score and information to be calculated. Additional assumptions are needed to ensure enough regularity for maximum likelihood estimators to be asymptotically normal, for example[1]. The variable X_t is assumed to be generated

[1]We do not pursue, here, abstract conditions for the MLE in the $GINAR(p)$ model to be asymptotically normal and efficient. This is a topic we leave for further research.

according to

$$X_t = \alpha_1 \diamond X_{t-1} + \alpha_2 \diamond X_{t-2} + \cdots + \alpha_p \diamond X_{t-p} + \varepsilon_t \tag{1}$$

where, conditional on X_{t-k}, $\alpha_k \diamond X_{t-k}$ is an integer-valued random variable (random operator) with parameter α_k[2]. The variables $\alpha_k \diamond X_{t-k}$, $k \in \{1, ..., p\}$, conditional on X_{t-k}, $k \in \{1, ..., p\}$, are mutually independent. The operator thus delivers an integer value and dependence in $\{X_t\}$ is induced via the conditioning variables X_{t-k}, $k \in \{1, ..., p\}$. The operator $\alpha_k \diamond X_{t-k}$ may correspond to Binomial thinning and with ε_t a Poisson variable this gives rise to the standard $INAR(p)$-P model; when $p = 1$, X_t has a Poisson distribution. Other possibilities are that conditional on X_{t-k}, $\alpha_k \diamond X_{t-k}$ is Beta-Binomial while ε_t is Negative Binomial; when $p = 1$ this will ensure that X_t is also Negative Binomial. For a general treatment of such operators see Joe (1996). The conditional probability density function of $\alpha_k \diamond X_{t-k}$ given X_{t-k}, with respect to the counting measure υ, is written

$$f(s_k | X_{t-k}; \alpha_k) \tag{2}$$

while that of ε_t is

$$g(\varepsilon; \boldsymbol{\lambda}). \tag{3}$$

In the calculations required for the score and the information matrix we assume that these densities satisfy

$$\begin{aligned} \frac{\partial f(s_k | X_{t-k}; \alpha_k)}{\partial \alpha_k} &= \tau(s_k; X_{t-k}, \alpha_k) f(s_k | X_{t-k}; \alpha_k) \\ \frac{\partial g(\varepsilon; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= \boldsymbol{\gamma}(\varepsilon; \boldsymbol{\lambda}) g(\varepsilon; \boldsymbol{\lambda}) \end{aligned} \tag{4}$$

where $\tau(\cdot)$ and the vector function $\boldsymbol{\gamma}(\cdot)$ are differentiable with respect to the parameters.

The next two sub-sections look at a likelihood based analysis of the $GINAR(p)$ model. We condition on the first p observations[3]. The first sub-section looks at the conditional likelihood while the second treats the score and information quantities.

[2] In fact, α_k may be a vector but we stick to the simpler scalar notation.

[3] The full likelihood may, however, be computed under stationarity as $P(X_1, .., X_p)$ may be calculated numerically from the conditional distributions. See Bu (2006) for details.

2.1. The Conditional Likelihood

Conditioning on the first p observations leads to a simple form of the likelihood viz.

$$L(\alpha_1, \ldots, \alpha_p, \boldsymbol{\lambda}) = \prod_{t=p+1}^{T} P(X_t | X_{t-1}, \ldots, X_{t-p}) \tag{5}$$

and so knowledge of the transition probabilities is sufficient for its construction. Theorem 2.1 below shows how these conditional probabilities may be calculated by a simple recursive mechanism. The idea is to regard X_t as the convolution of $\alpha_1 \diamond X_{t-1}$ and $\alpha_2 \diamond X_{t-2} + \cdots + \alpha_p \diamond X_{t-p} + \varepsilon_t$, which are by definition mutually independent given the p observed lags. Then $\alpha_2 \diamond X_{t-2} + \cdots + \alpha_p \diamond X_{t-p} + \varepsilon_t$ may be thought of the the convolution between $\alpha_2 \diamond X_{t-2}$ and $\alpha_3 \diamond X_{t-3} + \cdots + \alpha_p \diamond X_{t-p} + \varepsilon_t$ and this leads to an obvious recursion.

Theorem 2.1. *In the model (1)*

$$\begin{aligned} & P(X_t | X_{t-1}, \ldots, X_{t-p}) \\ = & \int f(s_1 | X_{t-1}) P(X_t - s_1 | X_{t-2}, \ldots, X_{t-p}) d\upsilon(s_1) \end{aligned} \tag{6}$$

where the starting value is given by

$$P\left(X_t - \sum_{k=1}^{p-1} s_k \middle| X_{t-p} \right) = \int f\left(s_p | X_{t-p}\right) g\left(X_t - \sum_{k=1}^{p-1} s_k - s_p \right) d\upsilon(s_p). \tag{7}$$

Theorem 2.1 allows the conditional likelihood of the $GINAR(p)$ model (1) to be calculated for any innovations sequence $\{\varepsilon_t\}$ and thinning variables via (5). In addition to facilitating computation of the (conditional) likelihood, the recursions of Theorem 2.1 are also very useful in computing derivatives and hence the score and information quantities.

2.2. The Score and Information

As in FM it proves convenient to express the score function in terms of certain conditional expectations. The following theorems extend the $GINAR(1)$ results of Freeland (1998).

Theorem 2.2. *Let $\dot{\ell}_{\alpha_k}$ denote the score with respect to α_k for $k \in \{1, ..., p\}$ and $\dot{\ell}_{\boldsymbol{\lambda}}$ the score with respect to the vector $\boldsymbol{\lambda}$. Denote by $E_t[\cdot]$ the conditional expectation with respect to the sigma field, $\Im_t = \sigma\left(X_t, X_{t-1}, ..., X_{t-p}\right)$. Assume the density functions f and g in (2) and (3) satisfy (4). Then for the model (1)*

$$\dot{\ell}_{\alpha_k} = \sum_{t=p+1}^{T} E_t \left[\tau \left(\alpha_k \diamond X_{t-k}\right)\right]$$

and

$$\dot{\ell}_{\boldsymbol{\lambda}} = \sum_{t=p+1}^{T} E_t[\boldsymbol{\gamma} \left(\varepsilon_t\right)]$$

where $\tau \left(\alpha_k \diamond X_{t-k}\right) = \tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k\right)$ and $\boldsymbol{\gamma} \left(\varepsilon_t\right) = \boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda}\right)$.

It is important to note that the time t expectations are different from those calculated at time $t-1$.

The information matrix can also be expressed in a similar way in terms of conditional expectations.

Theorem 2.3. *Let $\ddot{\ell}_{ab}$ denote the second derivatives of the log-likelihood with respect to a and b and let τ_{α_k} denote the derivative of the function τ with respect to α_k. The matrix $\boldsymbol{\gamma}_{\boldsymbol{\lambda}}$ is defined as the derivative of the vector function $\boldsymbol{\gamma}$ with respect to the vector $\boldsymbol{\lambda}$. Under the conditions of Theorem 2.2 the following results hold for the model (1):*

$$\ddot{\ell}_{\alpha_k \alpha_k} = \sum_{t=p+1}^{T} \left\{E_t \left[\tau_{\alpha_k} \left(\alpha_k \diamond X_{t-k}\right)\right] + Var_t \left[\tau \left(\alpha_k \diamond X_{t-k}\right)\right]\right\},$$

$$\ddot{\ell}_{\alpha_m \alpha_n} = \sum_{t=p+1}^{T} Cov_t \left[\tau \left(\alpha_m \diamond X_{t-m}\right), \tau \left(\alpha_n \diamond X_{t-n}\right)\right],$$

$$\ddot{\ell}_{\alpha_k \boldsymbol{\lambda}} = \sum_{t=p+1}^{T} Cov_t \left[\tau \left(\alpha_k \diamond X_{t-k}\right), \boldsymbol{\gamma} \left(\varepsilon_t\right)\right]$$

and

$$\ddot{\ell}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} = \sum_{t=p+1}^{T} \left\{E_t \left[\boldsymbol{\gamma}_{\boldsymbol{\lambda}} \left(\varepsilon_t; \boldsymbol{\lambda}\right)\right] + Var_t \left[\boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda}\right)\right]\right\}$$

where $\tau_{\alpha_k} \left(\alpha_k \diamond X_{t-k}\right) = \tau_{\alpha_k} \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k\right)$ and $\boldsymbol{\gamma}_{\boldsymbol{\lambda}} \left(\varepsilon_t\right) = \boldsymbol{\gamma}_{\boldsymbol{\lambda}} \left(\varepsilon_t; \boldsymbol{\lambda}\right)$.

In the remainder of the paper these results are utilised when the thinning is Binomial and the arrivals are Poisson.

3. The $INAR(p)$-P Model

In the spirit of Du and Li (1991) (DL) we define the $INAR(p)$-P to be

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t \tag{8}$$

where the innovation process $\{\varepsilon_t\}$ is an i.i.d Poisson process. The innovations are assumed to be independent of all thinning operations $\alpha_k \circ X_{t-k}$, $k \in \{1, ..., p\}$. Conditional on X_{t-k}, $k \in \{1, ..., p\}$, the thinning operators are Binomial, defined as

$$\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} \mathrm{B}_{i,k}$$

where each collection $\{\mathrm{B}_{i,k}, i = 1, ..., X_{t-k}\}$ consists of independently distributed Bernoulli random variables with thinning parameter α_k and the collections are mutually independent $k \in \{1, ..., p\}$. The case where $p = 1$ is known as Poisson autoregression since in this case the marginal stationary distribution of X_t is also Poisson. When $p > 1$ it can be shown that the unconditional mean of X_t and the unconditional variance of X_t are generally not equal so that the marginal stationary distribution of X_t is no longer Poisson even though the innovations are. DL show that, for $\alpha_k \in [0, 1)$, (8) is stationary as long as $\sum_{k=1}^{p} a_k < 1$ and that the correlation properties of this process are identical to the linear Gaussian $AR(p)$ model. Dion et al. (1995) show that the $INAR(p)$ process may be generally viewed as a special multitype branching process with immigration.

The Alzaid and Al-Osh (1990) specification of the $INAR(p)$ process differs from that of DL in that it employs an alternative assumption that the conditional distribution of the $(\alpha_1 \circ X_{t-p}, \alpha_2 \circ X_{t-p}, \ldots, \alpha_p \circ X_{t-p})'$ given X_{t-p} is multinomial with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_p, X_{t-p})$. The statistical properties of the Alzaid and Al-Osh (1990) model are very different from that of DL and the model is much less tractable. In this study, we confine ourselves to the case where the thinning operators are conditionally independent.

The following proposition gives an explicit expression for the conditional probabilities of Theorem 2.1 in the case of Poisson innovations and Binomial thinning and it follows directly from Theorem 2.1 above by substitution.

Proposition 3.1. *For the $INAR(p)$-P model,*

$$
\begin{aligned}
& P(X_t|X_{t-1},\ldots,X_{t-p}) \\
= & \sum_{i_1=0}^{\min(X_{t-1},X_t)} \binom{X_{t-1}}{i_1} \alpha_1^{i_1}(1-\alpha_1)^{X_{t-1}-i_1} \sum_{i_2=0}^{\min[X_{t-2},X_t-i_1]} \binom{X_{t-2}}{i_2} \alpha_2^{i_2}(1-\alpha_2)^{X_{t-2}-i_2} \\
& \cdots \sum_{i_p=0}^{\min[X_{t-p},X_t-(i_1+\cdots+i_{p-1})]} \binom{X_{t-p}}{i_p} \alpha_p^{i_p}(1-\alpha_p)^{X_{t-p}-i_p} \frac{e^{-\lambda}\lambda^{X_t-(i_1+\cdots+i_p)}}{[X_t-(i_1+\cdots+i_p)]!}. \quad (9)
\end{aligned}
$$

From the transition probabilities the likelihood may be calculated. Under the stationarity assumption $\sum_{k=1}^{p} a_k < 1$, which we now assume, conditioning on the initial observations will have little effect when the sample size is reasonably large. Further, this simplification will not affect the ARE comparisons in Section 4.

In the case where the thinning is Binomial the τ function has the explicit form

$$\tau\left(s_k; X_{t-k}, \alpha_k\right) = \frac{s_k}{\alpha_k(1-\alpha_k)} - \frac{\alpha_k X_{t-k}}{\alpha_k(1-\alpha_k)}$$

and the score functions for α_k are

$$\dot{\ell}_{\alpha_k} = \frac{1}{\alpha_k(1-\alpha_k)} \sum_{t=p+1}^{T} \left\{E_t[\alpha_k \circ X_{t-k}] - E_{t-1}[\alpha_k \circ X_{t-k}]\right\}.$$

Thus, this score measures the incremental information contribution of the thinning operator. When the arrivals are Poisson the γ function takes the form

$$\gamma\left(\varepsilon;\lambda\right) = \frac{\varepsilon}{\lambda} - 1$$

and the score for λ is

$$\dot{\ell}_{\lambda} = \frac{1}{\lambda} \sum_{t=p+1}^{T} \left\{E_t[\varepsilon_t] - E_{t-1}[\varepsilon_t]\right\}$$

which also has informational interpretation for the innovations process[4]. The following proposition notes the information quantities for the Binomial-Poisson case.

[4]Note at time $t-1$, $E_{t-1}\left[\alpha_k \circ X_{t-k}\right] = \alpha_k X_{t-k}$ and $E_{t-1}\left[\varepsilon_t\right] = \lambda$.

Proposition 3.2. *Under the conditions of Theorem 2.2 the following results hold for the $INAR(p)$-P model:*

$$\begin{aligned}\ddot{\ell}_{\alpha_k\alpha_k} &= \frac{1}{\alpha_k^2\left(1-\alpha_k\right)^2}\sum_{t=p+1}^{T}\left\{(2\alpha_k-1)E_t[\alpha_k\circ X_{t-k}]\right.\\ &\quad \left.+Var_t[\alpha_k\circ X_{t-k}]-\alpha_k E_{t-1}[\alpha_k\circ X_{t-k}]\right\},\end{aligned}$$

$$\ddot{\ell}_{\alpha_m\alpha_n}=\frac{1}{\alpha_m\alpha_n(1-\alpha_m)(1-\alpha_n)}\sum_{t=p+1}^{T}Cov_t[\alpha_m\circ X_{t-m},\alpha_n\circ X_{t-n}],$$

$$\ddot{\ell}_{\alpha_k\lambda}=\frac{1}{\lambda\alpha_k(1-\alpha_k)}\sum_{t=p+1}^{T}Cov_t[\alpha_k\circ X_{t-k},\varepsilon_t]$$

and

$$\ddot{\ell}_{\lambda\lambda}=\frac{1}{\lambda^2}\sum_{t=p+1}^{T}\left\{Var_t[\varepsilon_t]-E_t[\varepsilon_t]\right\}.$$

These representations clearly show that the scores and information implied by the $INAR(p)$-P model can be decomposed into quantities associated with each component of the model. For example, the expression $\ddot{\ell}_{\lambda\lambda}$ reflects the Poisson mean-variance relationship given the additional information available at time t and the off-diagonal elements reflect the covariances between the unobserved components of the model.

In addition to enhancing the interpretation of the model these conditional expectations are also an important computational tool. For example,

$$\begin{aligned}E_t\left[\alpha_k\circ X_{t-k}\right] &= \frac{\alpha_k X_{t-k}P(X_t-1|X_{t-1},\ldots,X_{t-k}-1,\ldots,X_{t-p})}{P(X_t|X_{t-1},\ldots,X_{t-p})},\\ E_t[\varepsilon_t] &= \frac{\lambda P(X_t-1|X_{t-1},\ldots,X_{t-p})}{P(X_t|X_{t-1},\ldots,X_{t-p})}\end{aligned}$$

and the conditional probabilities required may be computed by (9) above and the expressions given in Appendix B.

The proof of the following theorem, which establishes the asymptotic normality of the ML estimator, is given in Appendix A.

Theorem 3.3. *Let $\boldsymbol{\theta} = (\alpha_1, ..., \alpha_p, \lambda)'$ denote the parameter vector for the stationary $INAR(p)$-P model (8). The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ has the following asymptotic distribution:*

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{i}^{-1})$$

where the matrix $\mathbf{i}$ *is the Fisher information per observation, i.e. the expectation of the second derivatives as given in Proposition 3.2.*

The parameter estimates for the model can be found using Newton-Raphson type iterative procedures. Standard errors of the estimates are readily available from the observed Fisher information matrix. Alternatively, if the time series is comprised of low counts, the expected Fisher Information can also be calculated numerically using the results in Proposition 3.2. See Section 4.1 for details.

4. Comparison of Methods

In this section, we compare the ML with the CLS method of Klimko and Nelson (1978). The CLS estimator (CLSE) $\widehat{\boldsymbol{\theta}}_{CLS}$ is strongly consistent and has the following asymptotic distribution (see DL)

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_{CLS} - \boldsymbol{\theta}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{j}^{-1})$$

where $\mathbf{j}$ is the Godambe information matrix given by

$$\mathbf{j} = \mathbf{S}\mathbf{V}^{-1}\mathbf{S}$$

with

$$\mathbf{S} = E\left[\frac{\partial g_t(\boldsymbol{\theta}, \Im_{t-1})}{\partial \boldsymbol{\theta}} \frac{\partial g_t(\boldsymbol{\theta}, \Im_{t-1})}{\partial \boldsymbol{\theta}'}\right], \tag{10}$$

$$\mathbf{V} = E\left[u_t^2(\boldsymbol{\theta}) \frac{\partial g_t(\boldsymbol{\theta}, \Im_{t-1})}{\partial \boldsymbol{\theta}} \frac{\partial g_t(\boldsymbol{\theta}, \Im_{t-1})}{\partial \boldsymbol{\theta}'}\right] \tag{11}$$

and

$$\begin{aligned} u_t(\boldsymbol{\theta}) &= X_t - g_t(\boldsymbol{\theta}, \Im_{t-1}) \\ &= X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \cdots - \alpha_p X_{t-p} - \lambda. \end{aligned}$$

In the $INAR(p)$ model, CLS estimation parallels OLS in traditional AR models. Of course, CLS only enforces the conditional mean restriction embodied in $u_t(\boldsymbol{\theta})$ and does not incorporate other conditional moment restrictions e.g. it does not take account of the conditional heteroscedasticity in the model. Thus, we may expect a certain loss of efficiency in comparison with ML when the model is true. We compare ML and CLS by evaluating the asymptotic relative efficiency (ARE) between the two estimators. The ARE between estimators is defined as the ratio of their asymptotic variances (see Cox and Hinkley (1974)). Let $\widehat{\boldsymbol{\theta}}$ be an estimate of $\boldsymbol{\theta}$ and denote by i_{kk}^{-1} the (k,k) element of $\mathbf{i}^{-1}$, the inverse of the Fisher information matrix. Similarly, let j_{kk}^{-1} be the (k,k) element of $\mathbf{j}^{-1}$, which is the inverse of the Godambe information matrix. The ARE for the k^{th} component of $\widehat{\boldsymbol{\theta}}$ is then defined as

$$ARE(\widehat{\theta}_{kk}) = \frac{i_{kk}^{-1}}{j_{kk}^{-1}}.$$

Clearly, in this setup, an ARE less than unity would suggest better efficiency for the MLE. Notice that there are no simulations involved in this comparison and the sample size is infinitely large. Furthermore, the comparison is between ML and CLS, i.e. conditioning on the initial observations has a negligible asymptotic effect here.

4.1. The $INAR(2)$-P Specification

In our comparison, we entertain the $INAR(2)$-P specification

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \varepsilon_t$$

where ε_t has a Poisson distribution with mean equal to λ. For ML, the expected Fisher information matrix can be written as

$$\mathbf{i} = \left(-E\left[\frac{\partial^2 \ln P(X_t | X_{t-1}, X_{t-2})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \right)^{-1} \tag{12}$$

where $P(X_t|X_{t-1}, X_{t-2})$ is the probability of X_t conditioned on X_{t-1} and X_{t-2}. Following Proposition 3.1, this conditional probability is given by

$$
\begin{aligned}
& P(X_t|X_{t-1},X_{t-2}) \\
= & \sum_{i=0}^{\min(X_{t-1},X_t)} \left\{ \binom{X_{t-1}}{i} \alpha_1^i (1-\alpha_1)^{X_{t-1}-i} \right. \\
& \left. \sum_{j=0}^{\min(X_{t-2},X_t-i)} \binom{X_{t-2}}{j} \alpha_2^j (1-\alpha_2)^{X_{t-2}-j} \frac{e^{-\lambda}\lambda^{X_t-i-j}}{(X_t-i-j)!} \right\} . \quad (13)
\end{aligned}
$$

By Proposition 3.2,

$$
\frac{\partial^2 \ln P(X_t|X_{t-1},X_{t-2})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \ddot{\ell}_{\alpha_1\alpha_1} & \ddot{\ell}_{\alpha_1\alpha_2} & \ddot{\ell}_{\alpha_1\lambda} \\ \ddot{\ell}_{\alpha_1\alpha_2} & \ddot{\ell}_{\alpha_2\alpha_2} & \ddot{\ell}_{\alpha_2\lambda} \\ \ddot{\ell}_{\alpha_1\lambda} & \ddot{\ell}_{\alpha_2\lambda} & \ddot{\ell}_{\lambda\lambda} \end{bmatrix}
$$

where each element in this information matrix can be calculated as specified in Appendix B. The expectation in (12) is calculated numerically. Specifically, we select a large enough positive integer value M such that the probability of a count larger than M is negligible. Then, for the $INAR(2)$-P model, there are $(M+1)^3$ possible outcomes of the joint observation of $\{X_t, X_{t-1}, X_{t-2}\}$ to sum over for each element of the Fisher information[5]. For example, summing over all $(M+1)^3$ possible values of $\{X_t, X_{t-1}, X_{t-2}\}$,

$$
\begin{aligned}
E[\ddot{\ell}_{\lambda\lambda}] = & \sum_{all\ \{X_t,X_{t-1},X_{t-2}\}} P(X_t, X_{t-1}, X_{t-2}) \\
& \times \left\{ \frac{P(X_t-2|X_{t-1},X_{t-2})}{P(X_t|X_{t-1},X_{t-2})} - \left[\frac{P(X_t-1|X_{t-1},X_{t-2})}{P(X_t|X_{t-1},X_{t-2})} \right]^2 \right\}
\end{aligned}
$$

where $P(X_t, X_{t-1}, X_{t-2})$ is the joint probability of $\{X_t, X_{t-1}, X_{t-2}\}$, which is also calculated numerically using the conditional probability function in (13). Details of transforming the conditional probabilities into the joint probability $P(X_t, X_{t-1}, X_{t-2})$ for stationary processes are given in Bu (2006). The expectation in both (10) and (11) are evaluated numerically in the same way as for the MLE case. GAUSS programs are available on request to perform these calculations.

[5] If $M = 6$, for instance, there are 343 possible outcomes of joint observation of $\{X_t, X_{t-1}, X_{t-2}\}$. They are $\{0,0,0\}$, $\{0,0,1\}$,..., and $\{6,6,6\}$.

4.2. ARE

We calculate and examine the ARE of the two estimators for a range of different parameter values. To ensure that the processes examined are stationary and non-degenerate, the sum of the two thinning parameters, α_1 and α_2, is confined within the range of $[0.10, 0.90]$ and, for each of the two thinning parameters, a sequence of different values ranging from 0.10 to 0.80, on a grid of 0.10, is entertained. All possible combinations of α_1 and α_2 are examined. We also try three different values of λ, (0.5, 1, and 2) to reflect varied arrival rates. However, we found that our qualitative conclusions are not affected by the choice of λ. Thus, for economy of space we only present results for the case where $\lambda = 1$. All unreported results are available upon request.

[Table 1]

Table 1 shows the ARE ratios for the parameters, $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, and $\widehat{\lambda}$, respectively. As expected, our results confirm that the MLE is asymptotically more efficient than the CLSE for all three parameters, since all the ARE ratios are less than unity. Generally, it is true that more substantial efficiency gains can be obtained from using the ML as the process becomes more persistent (higher values of α_1 or α_2, or both). Specifically, it can be seen from Table 1, Panel 1, that, for a given value of α_2, the ARE of $\widehat{\alpha}_1$ decreases as the value of α_1 increases with the largest advantage for ML occurring when α_1 is large and α_2 is small. Table 1, Panel 2, shows that the ARE of $\widehat{\alpha}_2$ is largest when either α_1 is large and α_2 is small or α_1 is small and α_2 is large. The third Panel of Table 1 confirms that substantial gains are obtained for estimating λ from persistent processes, especially if either $\widehat{\alpha}_1$ or $\widehat{\alpha}_2$ approaches unity. But it is interesting to note that, unlike the previous two cases, the ARE of $\widehat{\lambda}$ is slightly more sensitive to the scale of α_2 and that the gains are never as large as those available for estimating either of the α's.

4.3. Monte Carlo Results

In order to compare the relative performance of the estimators in small samples, we carry out Monte Carlo experiments to examine the finite sample bias and mean squared error (MSE) of alternative estimators[6]. To achieve this, we generate artificial time series of counts based on the $INAR(2)$-P model. As before, values

[6]We also included the Yule-Walker estimator in the simulations but found its performance very similar to that of CLS and hence these results are not reported.

of α_1 and α_2 as well as their sum, $(\alpha_1 + \alpha_2)$, are constrained so that each case under study is stationary and non-degenerate. The study is based on 1000 replications. For each replication, we estimate the model parameters using alternative estimators and calculate the bias and MSE of parameter estimates. Our simulation experiments are performed for sample size $T = 100$ and 500. As in the ARE calculations the qualitative results do not depend on the value of λ and so we report the $\lambda = 1$ case only.

[Table 2]
[Table 3]

The bias results are reported in Tables 2 and 3. It can be seen that, except for a few cases where the ML of $\widehat{\alpha}_1$ is biased up, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are both biased down and $\widehat{\lambda}$ is biased up. This inverse relationship is to be expected because, for a fixed marginal mean of the series X_t, decreasing α_1 and α_2 corresponds to increasing λ. The relationship between the bias in $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$, however, is less evident from the table. But a closer examination of all cases studied, including those unreported, also reveals a negative correlation, despite the fact that both are biased down. This inverse relationship is also expected for similar reason. That is, for a fixed marginal mean of X_t and λ, a large α_1 corresponds to a small α_2, and vice versa. With respect to sample size, Table 3 suggests that the bias of both the CLS and ML estimates is inversely related to the sample size with a minor exception in the case of $\hat{\alpha}_1$ for very small values of α_1 and α_2. Unless the parameters are in the vicinity of the nonstationary region, the bias of the MLE is less that of CLS for both $\hat{\alpha}_1$ and $\hat{\alpha}_2$. The MLE of λ always dominates in terms of bias. These results hold even at $T = 500$. This suggests that there is a gain in using the ML over CLS in terms of bias except when close to the nonstationary region.

[Table 4]
[Table 5]

The corresponding MSE results are given in Tables 4 and 5. In the case of $\hat{\alpha}_1$ Table 4, Panel 1, shows that for small values of α_1 the MSE of ML is greater than that of CLS for $T = 100$. The corresponding phenomenon holds for $\hat{\alpha}_2$ as seen in Panel 2 of Table 4. The MSE of the MLE of $\hat{\lambda}$ is less that that of CLS (Table 4, Panel 3). Table 5 shows that these anomalies disappear at $T = 500$ and the MSE of the MLE of all parameters is smallest. Certainly, for larger sample sizes there is a gain in terms of MSE in using the ML method.

5. Conclusion

In this paper, we present a framework for maximum likelihood estimation of $GINAR(p)$ processes based on a recursive representation of the transition probabilities. Using the resulting likelihood, we derive the score function and the Fisher information matrix for the model, which form the basis for conditional maximum likelihood estimation and inference. As in FM, we go on to represent all elements of the Fisher information matrix in terms of time t conditional moments of model components. Using the $INAR(2)$-P specification, we investigate the asymptotic gain of implementing the ML method over the commonly used CLS method by calculating the ARE ratio between the two estimators. Our results confirm that the proposed MLE is asymptotically more efficient than the CLSE and the efficiency gain is most substantial for persistent processes. A Monte Carlo study suggests that there are often small sample gains in terms of bias and MSE to be had.

The proposed maximum likelihood framework also allows for various types of likelihood-based statistical inferences. For instance, given the score functions and elements of the Fisher information matrix, it is possible to test for model adequacy using the information matrix test proposed by McCabe and Leybourne (2000). Moreover, the transition probability function for the $INAR(p)$-P process provides a basis for coherent forecasting. These and other issues involved in model selection are examined in Bu and McCabe (2008).

References

Al-Osh M.A. and Alzaid, A.A. (1987) First-order integer valued autoregressive (INAR(1)) process. Journal of Time Series Analysis 8, 261-275.

Alzaid, A.A. and Al-Osh, M.A. (1990) An integer-valued pth-order autoregressive structure (INAR(p)) process. Journal of Applied Probability 27, 314-323.

Azzalini, A. (1983) Maximum likelihood estimation of order m stationary stochastic processes. Biometrika, 70, 381-387.

Bu, R. (2006) Essays in financial econometrics and time series analysis. Ph.D Thesis, University of Liverpool, UK.

Bu, R. and McCabe, B.P.M. (2008) Model selection, estimation and forecasting in INAR(p) models: a likelihood based Markov Chain approach. International Journal of Forecasting, 24, 151-162.

Cox D.R. and Hinkley, D. (1974) Theoretical statistics, Chapman and Hall, London.

Davidson, J. (1994) Stochastic limit theory, Oxford University Press.

Dion, J.P., Gauthier, G., and Latour, A. (1995) Branching processes with immigration and integer-valued time series. Serdica, 21, 123-136.

Drost, F.C., Van den Akker, R., and Werker, B.J.M. (2008) Local Asymptotic Normality and efficient estimation for INAR(p) models. Journal of Time Series Analysis, Forthcoming.

Du, J. and Li, Y. (1991) The integer-valued autoregressive (INAR(p)) model. Journal of Time Series Analysis 12, 129-142.

Freeland, R.K. (1998) Statistical analysis of discrete time series with applications to the analysis of workers compensation claims data. Ph.D. Thesis, The University of British Columbia, Canada.

Freeland, R.K. and McCabe, B.P.M. (2004) Analysis of low count time series by Poisson autoregression, Journal of Time Series Analysis, 25, 701-722.

Joe, H. (1996) Time series models with univariate margins in the convolution-closed infinitely divisible class. Journal of Applied Probability 33, 664-677.

Jung, R.C. and Tremayne, A.R. (2006) Coherent forecasting in integer time series models. International Journal of Forecasting, 22, 223-238.

Klimko, L.A. and Nelson, P.I. (1978) On conditional least squares estimation for stochastic processes. Annals of Statistics 6, 629-642.

McCabe, B.P.M. and Leybourne, S.J. (2000) A general method of testing for random parameter variation in statistical models in *Innovations in Multivariate Statistical Analysis: a Festschrift for Heinz Neudecker*, eds. Heijmans, R.D.H, D.S.G. Pollock and A. Satorra, R.D.H, 75-85, Kluwer.

McKenzie, E. (1988) Some ARMA models for dependent sequences of Poison counts. Advances in Applied Probability 20, 822-835.

Table 1: Asymptotic Relative Efficiency for the $INAR(2)$-P Model ($\lambda = 1$)

		α_2							
	α_1	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
	0.10	0.9622	0.9573	0.9476	0.9362	0.9228	0.9030	0.8680	0.7750
	0.20	0.8945	0.9013	0.9051	0.9047	0.8953	0.8635	0.7627	
	0.30	0.8071	0.8305	0.8531	0.8709	0.8724	0.8182		
ARE(α_1)	0.40	0.6995	0.7411	0.7885	0.8364	0.8620			
	0.50	0.5678	0.6240	0.6969	0.7854				
	0.60	0.4153	0.4728	0.5541					
	0.70	0.2564	0.2945						
	0.80	0.1169							

		α_2							
	α_1	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
	0.10	0.9622	0.8933	0.8067	0.7110	0.6088	0.4994	0.3827	0.2572
	0.20	0.9579	0.9015	0.8264	0.7397	0.6432	0.5362	0.4157	
	0.30	0.9452	0.9091	0.8538	0.7853	0.7047	0.6066		
ARE(α_2)	0.40	0.9122	0.9008	0.8734	0.8370	0.7894			
	0.50	0.8412	0.8503	0.8517	0.8544				
	0.60	0.7159	0.7229	0.7205					
	0.70	0.5300	0.4876						
	0.80	0.2943							

		α_2							
	α_1	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
	0.10	0.9546	0.9074	0.8479	0.7850	0.7239	0.6668	0.6239	0.6232
	0.20	0.9091	0.8507	0.7859	0.7206	0.6607	0.6154	0.6120	
	0.30	0.8597	0.7923	0.7211	0.6521	0.5966	0.5804		
ARE(λ)	0.40	0.8135	0.7371	0.6590	0.5898	0.5552			
	0.50	0.7733	0.6901	0.6092	0.5557				
	0.60	0.7444	0.6589	0.5921					
	0.70	0.7307	0.6562						
	0.80	0.7385							

Table 2: Bias Results for the $INAR(2)$-P Model ($\lambda = 1$) for Sample Size $T = 100$

	CLS					CML				
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	-0.0032	-0.0057	-0.0118	-0.0168	0.10	0.0019	-0.0018	-0.0080	-0.0177
Bias(α_1)	0.30	-0.0163	-0.0200	-0.0206		0.30	-0.0066	-0.0123	-0.0118	
	0.50	-0.0219	-0.0190			0.50	-0.0027	-0.0017		
	0.70	-0.0344				0.70	-0.0021			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	-0.0097	-0.0340	-0.0436	-0.0491	0.10	-0.0066	-0.0214	-0.0214	-0.0099
Bias(α_2)	0.30	-0.0133	-0.0377	-0.0491		0.30	-0.0104	-0.0264	-0.0231	
	0.50	-0.0135	-0.0374			0.50	-0.0142	-0.0291		
	0.70	-0.0111				0.70	-0.0233			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0214	0.0562	0.1267	0.3143	0.10	0.0110	0.0283	0.0616	0.1193
Bias(λ)	0.30	0.0485	0.1263	0.3206		0.30	0.0274	0.0791	0.1451	
	0.50	0.0832	0.2700			0.50	0.0362	0.1423		
	0.70	0.2080				0.70	0.1068			

Table 3: Bias Results for the $INAR(2)$-P Model ($\lambda = 1$) for Sample Size $T = 500$

	CLS					CML				
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	-0.0028	-0.0008	-0.0043	-0.0028	0.10	-0.0022	0.0007	-0.0038	-0.0034
Bias(α_1)	0.30	-0.0030	-0.0033	-0.0042		0.30	-0.0011	-0.0017	-0.0018	
	0.50	-0.0031	-0.0036			0.50	-0.0001	0.0007		
	0.70	-0.0035				0.70	0.0001			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	-0.0034	-0.0066	-0.0093	-0.0112	0.10	-0.0029	-0.0035	-0.0035	-0.0024
Bias(α_2)	0.30	-0.0043	-0.0058	-0.0089		0.30	-0.0033	-0.0035	-0.0042	
	0.50	-0.0069	-0.0072			0.50	-0.0062	-0.0056		
	0.70	-0.0055				0.70	-0.0061			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0040	0.0131	0.0338	0.0682	0.10	0.0027	0.0052	0.0180	0.0272
Bias(λ)	0.30	0.0091	0.0226	0.0632		0.30	0.0042	0.0128	0.0280	
	0.50	0.0233	0.0533			0.50	0.0143	0.0231		
	0.70	0.0428				0.70	0.0275			

Table 4: MSE Results for the $INAR(2)$-P Model ($\lambda = 1$) for Sample Size $T = 100$

	CLS					CML				
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0073	0.0074	0.0061	0.0053	0.10	0.0080	0.0078	0.0065	0.0051
MSE(α_1)	0.30	0.0115	0.0122	0.0095		0.30	0.0103	0.0112	0.0101	
	0.50	0.0101	0.0119			0.50	0.0063	0.0085		
	0.70	0.0102				0.70	0.0030			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0074	0.0113	0.0111	0.0087	0.10	0.0077	0.0099	0.0071	0.0028
MSE(α_2)	0.30	0.0071	0.0118	0.0112		0.30	0.0075	0.0113	0.0088	
	0.50	0.0067	0.0119			0.50	0.0068	0.0123		
	0.70	0.0070				0.70	0.0052			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0330	0.0487	0.1015	0.3166	0.10	0.0329	0.0417	0.0689	0.1302
MSE(λ)	0.30	0.0496	0.0995	0.3388		0.30	0.0427	0.0739	0.1557	
	0.50	0.0708	0.3013			0.50	0.0503	0.1608		
	0.70	0.2059				0.70	0.1100			

Table 5: MSE Results for the $INAR(2)$-P Model ($\lambda = 1$) for Sample Size $T = 500$

	CLS					CML				
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0021	0.0020	0.0017	0.0012	0.10	0.0021	0.0019	0.0016	0.0010
MSE(α_1)	0.30	0.0022	0.0020	0.0017		0.30	0.0019	0.0018	0.0016	
	0.50	0.0024	0.0021			0.50	0.0014	0.0015		
	0.70	0.0021				0.70	0.0006			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0022	0.0023	0.0020	0.0013	0.10	0.0022	0.0018	0.0012	0.0005
MSE(α_2)	0.30	0.0021	0.0021	0.0019		0.30	0.0020	0.0019	0.0014	
	0.50	0.0021	0.0020			0.50	0.0018	0.0018		
	0.70	0.0021				0.70	0.0012			
		α_2					α_2			
	α_1	0.10	0.30	0.50	0.70	α_1	0.10	0.30	0.50	0.70
	0.10	0.0080	0.0121	0.0198	0.0452	0.10	0.0077	0.0098	0.0136	0.0265
MSE(λ)	0.30	0.0102	0.0156	0.0379		0.30	0.0087	0.0113	0.0208	
	0.50	0.0145	0.0328			0.50	0.0112	0.0182		
	0.70	0.0287				0.70	0.0203			

Appendix A

The following straightforward result is used, often without comment, throughout the proofs. Let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a random vector and Y a random variable where $X_1, \ldots, X_p$ and Y are mutually independent. Denote their densities as $f_{\mathbf{X}}(\mathbf{x})$ and $f_Y(y)$. Let $Z = \mathbf{X}'\mathbf{1} + Y$ be the convolution of $X_1 + \ldots + X_p$, and Y, where $\mathbf{1}$ is a $p \times 1$ vector of ones. The conditional moments for $\phi(\mathbf{X}, Y)$ given Z are then

$$\begin{aligned} & E[\phi(\mathbf{X}, Y) | Z] \\ = \; & E[\phi(\mathbf{X}, Z - \mathbf{X}'\mathbf{1}) | Z] \\ = \; & \frac{\int \phi(\mathbf{x}, z - \mathbf{x}'\mathbf{1}) f_{\mathbf{X}}(\mathbf{x}) f_Y(z - \mathbf{x}'\mathbf{1}) d\mathbf{x}}{f_Z(z)}. \end{aligned} \tag{14}$$

We also use the following additional notation. The transition probability density function $P(X_t | X_{t-1}, \ldots, X_{t-p})$ is denoted by

$$h(X_t | X_{t-1}, \ldots, X_{t-p}; \alpha_1, \ldots, \alpha_p, \boldsymbol{\lambda}) = h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})$$

where $\mathbf{X}_{-p} = (X_{t-1}, \ldots, X_{t-p})'$, and $\boldsymbol{\lambda}$ may be a vector. To simplify integration with respect to a vector $\mathbf{s}$, we set $\mathbf{s} = (s_1, \ldots, s_p)'$ and $d\boldsymbol{\upsilon}_s = (d\upsilon(s_1), \cdots, d\upsilon(s_p))'$.

Proof (of Theorem 2.1) We regard X_t as the convolution of $\alpha_1 \diamond X_{t-1}$ and $Y = \alpha_2 \diamond X_{t-2} + \cdots + \alpha_p \diamond X_{t-p} + \varepsilon_t$, which are by definition mutually independent given the p observed lags. Thus, we can write

$$\begin{aligned} & h(X_t | X_{t-1}, \ldots, X_{t-p}; \alpha_1, \ldots, \alpha_p, \boldsymbol{\lambda}) \\ = \; & \int f(s_1 | X_{t-1}; \alpha_1) h_Y(X_t - s_1 | X_{t-2}, \ldots, X_{t-p}; \alpha_2, \ldots, \alpha_p, \boldsymbol{\lambda}) d\upsilon(s_1) \end{aligned}$$

where $h_Y(Y | X_{t-2}, \ldots, X_{t-p}; \alpha_2, \ldots, \alpha_p, \boldsymbol{\lambda})$ is the conditional probability density function of Y given observations $(X_{t-2}, \ldots, X_{t-p})$ and parameters $(\alpha_2, \ldots, \alpha_p, \boldsymbol{\lambda})$. It is important to note that the quantity $h_Y(Y | X_{t-2}, \ldots, X_{t-p}; \alpha_2, \ldots, \alpha_p, \boldsymbol{\lambda})$ can be evaluated using the expression of the transition probability density function for a $GINAR(p-1)$ process with parameters $(\alpha_2, \ldots, \alpha_p, \boldsymbol{\lambda})$. This is purely a computational device. We thus have the following recursive representation.

$$\begin{aligned} & h^{(p)}(X_t | X_{t-1}, \ldots, X_{t-p}; \alpha_1, \ldots, \alpha_p, \boldsymbol{\lambda}) \\ = \; & \int f(s_1 | X_{t-1}; \alpha_1) h^{(p-1)}(X_t - s_1 | X_{t-2}, \ldots, X_{t-p}; \alpha_2, \ldots, \alpha_p, \boldsymbol{\lambda}) d\upsilon(s_1). \end{aligned}$$

The superscript denotes that the conditional probability density function has the same expression as the transition probability of a $GINAR$ process with corresponding order. The recursion is initialised by

$$h^{(1)}\left(X_t - \sum_{k=1}^{p-1} s_k \middle| X_{t-p}; \alpha_p, \boldsymbol{\lambda}\right) = \int f\left(s_p | X_{t-p}, \alpha_p\right) g\left(X_t - \sum_{k=1}^{p} s_k; \boldsymbol{\lambda}\right) d\upsilon(s_p)$$

which is just the convolution of the $GINAR(1)$ model with arguments as specified. ■

Proof (of Theorem 2.2) The conditional log-likelihood function can be written as

$$\ln L(\boldsymbol{\theta}) = \sum_{t=p+1}^{T} \ln h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})$$

where

$$h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta}) = \int \left[\prod_{k=1}^{p} f(s_k | X_{t-k}; \alpha_k) g(X_t - \sum_{k=1}^{p} s_k; \boldsymbol{\lambda})\right] d\boldsymbol{\upsilon}_s.$$

We define

$$k\left(\mathbf{s}, X_t; \boldsymbol{\theta}\right) = \prod_{k=1}^{p} f(s_k | X_{t-k}; \alpha_k) g(X_t - \sum_{k=1}^{p} s_k; \boldsymbol{\lambda}).$$

Hence

$$h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta}) = \int k\left(\mathbf{s}, X_t; \boldsymbol{\theta}\right) d\boldsymbol{\upsilon}_s.$$

It follows that the corresponding score functions are given by

$$\dot{\ell}_{\alpha_k} = \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \alpha_k} = \sum_{t=p+1}^{T} \frac{\frac{\partial}{\partial \alpha_k} h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})}{h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})},$$

$$\dot{\ell}_{\boldsymbol{\lambda}} = \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\lambda}} = \sum_{t=p+1}^{T} \frac{\frac{\partial}{\partial \boldsymbol{\lambda}} h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})}{h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})}.$$

Under the conditions of Theorem 2.2

$$\begin{aligned} \frac{\partial f(s_k | X_{t-k}; \alpha_k)}{\partial \alpha_k} &= \tau\left(s_k; X_{t-k}, \alpha_k\right) f(s_k | X_{t-k}; \alpha_k), \\ \frac{\partial g(\varepsilon; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= \boldsymbol{\gamma}\left(\varepsilon; \boldsymbol{\lambda}\right) g(\varepsilon; \boldsymbol{\lambda}) \end{aligned}$$

and so

$$\begin{aligned}
\frac{\frac{\partial}{\partial\alpha_k}h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} &= \frac{\frac{\partial}{\partial\alpha_k}\left[\int k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)d\boldsymbol{\upsilon}_s\right]}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= \frac{\int \frac{\partial}{\partial\alpha_k}\left[k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)\right]d\boldsymbol{\upsilon}_s}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= \frac{\int \tau\left(s_k;X_{t-k},\alpha_k\right)k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)d\boldsymbol{\upsilon}_s}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= E_t\left[\tau\left(\alpha_k\diamond X_{t-k};X_{t-k},\alpha_k\right)\right].
\end{aligned}$$

Note that the last equality follows from the result on conditional expectations in (14). Similarly,

$$\begin{aligned}
\frac{\frac{\partial}{\partial\boldsymbol{\lambda}}h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} &= \frac{\frac{\partial}{\partial\boldsymbol{\lambda}}\left[\int k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)d\boldsymbol{\upsilon}_s\right]}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= \frac{\int \frac{\partial}{\partial\boldsymbol{\lambda}}\left[k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)\right]d\boldsymbol{\upsilon}_s}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= \frac{\int \boldsymbol{\gamma}\left(X_t-\sum_{k=1}^{p}s_k;\boldsymbol{\lambda}\right)k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)d\boldsymbol{\upsilon}_s}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= E_t\left[\boldsymbol{\gamma}\left(\varepsilon_t;\boldsymbol{\lambda}\right)\right].
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\dot{\ell}_{\alpha_k} &= \sum_{t=p+1}^{T} E_t\left[\tau\left(\alpha_k\diamond X_{t-k};X_{t-k},\alpha_k\right)\right], \\
\dot{\ell}_{\boldsymbol{\lambda}} &= \sum_{t=p+1}^{T} E_t\left[\boldsymbol{\gamma}\left(\varepsilon_t;\boldsymbol{\lambda}\right)\right].
\end{aligned}$$

■

Proof (of Theorem 2.3) Define scalar function $\tau_{\alpha_k}\left(s_k;X_{t-k},\alpha_k\right)$ and matrix function $\boldsymbol{\gamma}_{\boldsymbol{\lambda}}\left(\varepsilon;\boldsymbol{\lambda}\right)$ such that

$$\begin{aligned}
\tau_{\alpha_k}\left(s_k;X_{t-k},\alpha_k\right) &= \frac{\partial\tau\left(s_k;X_{t-k},\alpha_k\right)}{\partial\alpha_k}, \\
\boldsymbol{\gamma}_{\boldsymbol{\lambda}}\left(\varepsilon;\boldsymbol{\lambda}\right) &= \frac{\partial\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}'}.
\end{aligned}$$

Under the conditions of Theorem 2.2:

$$\begin{aligned}
\frac{\partial^2 f(s_k|X_{t-k};\alpha_k)}{\partial\alpha_k^2} &= \frac{\partial\left[\tau\left(s_k;X_{t-k},\alpha_k\right)f(s_k|X_{t-k};\alpha_k)\right]}{\partial\alpha_k} \\
&= \frac{\partial\tau\left(s_k;X_{t-k},\alpha_k\right)}{\partial\alpha_k}f(s_k|X_{t-k};\alpha_k) \\
&\quad + \left[\tau\left(s_k;X_{t-k},\alpha_k\right)\right]^2 f(s_k|X_{t-k};\alpha_k) \\
&= \left\{\tau_{\alpha_k}\left(s_k;X_k,\alpha_k\right) + \left[\tau\left(s_k;X_{t-k},\alpha_k\right)\right]^2\right\} f(s_k|X_{t-k};\alpha_k),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 g(\varepsilon;\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'} &= \frac{\partial\left[\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)g(\varepsilon;\boldsymbol{\lambda})\right]}{\partial\boldsymbol{\lambda}'} \\
&= \frac{\partial\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}'}g(\varepsilon;\boldsymbol{\lambda}) + \boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)\frac{\partial g(\varepsilon;\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}'} \\
&= \frac{\partial\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}'}g(\varepsilon;\boldsymbol{\lambda}) + \left[\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)'\right]g(\varepsilon;\boldsymbol{\lambda}) \\
&= \left\{\boldsymbol{\gamma}_{\boldsymbol{\lambda}}\left(\varepsilon;\boldsymbol{\lambda}\right) + \left[\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)\boldsymbol{\gamma}\left(\varepsilon;\boldsymbol{\lambda}\right)'\right]\right\}g(\varepsilon;\boldsymbol{\lambda}).
\end{aligned}$$

It then follows that

$$\begin{aligned}
\frac{\frac{\partial^2}{\partial\alpha_k^2}h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} &= \frac{\frac{\partial^2}{\partial\alpha_k^2}\left[\int k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)d\boldsymbol{\upsilon}_s\right]}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= \frac{\int\frac{\partial^2}{\partial\alpha_k^2}\left[k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)\right]d\boldsymbol{\upsilon}_s}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= \frac{\int\left\{\tau_{\alpha_k}\left(s_k;X_{t-k},\alpha_k\right) + \left[\tau\left(s_k;X_{t-k},\alpha_k\right)\right]^2\right\}k\left(\mathbf{s},X_t;\boldsymbol{\theta}\right)d\boldsymbol{\upsilon}_s}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \\
&= E_t\left[\tau_{\alpha_k}\left(\alpha_k\diamond X_{t-k};X_{t-k},\alpha_k\right) + \left[\tau\left(\alpha_k\diamond X_{t-k};X_{t-k},\alpha_k\right)\right]^2\right].
\end{aligned}$$

In exactly the same way we can show

$$\frac{\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} = E_t\left[\boldsymbol{\gamma}_{\boldsymbol{\lambda}}\left(\varepsilon_t;\boldsymbol{\lambda}\right) + \boldsymbol{\gamma}\left(\varepsilon_t;\boldsymbol{\lambda}\right)\boldsymbol{\gamma}\left(\varepsilon_t;\boldsymbol{\lambda}\right)'\right],$$

$$\frac{\frac{\partial^2}{\partial\alpha_k\partial\boldsymbol{\lambda}}h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} = E_t\left[\tau\left(\alpha_k\diamond X_{t-k};X_{t-k},\alpha_k\right)\boldsymbol{\gamma}\left(\varepsilon_t;\boldsymbol{\lambda}\right)\right],$$

$$\frac{\frac{\partial^2}{\partial\alpha_m\partial\alpha_n}h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} = E_t\left[\tau\left(\alpha_m\diamond X_{t-m};X_{t-m},\alpha_m\right)\tau\left(\alpha_n\diamond X_{t-n};X_{t-n},\alpha_n\right)\right].$$

Finally, the Fisher information can then be written as follows:

$$\begin{aligned}
\ddot{\ell}_{\alpha_k\alpha_k} &= \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \alpha_k^2} \\
&= \sum_{t=p+1}^{T} \left\{ \frac{\frac{\partial^2}{\partial \alpha_k^2} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} - \left[\frac{\frac{\partial}{\partial \alpha_k} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \right]^2 \right\} \\
&= \sum_{t=p+1}^{T} \left\{ E_t \left[\tau_{\alpha_k} \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) + \left[\tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) \right]^2 \right] \right. \\
&\quad \left. - \left(E_t \left[\tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) \right] \right)^2 \right\} \\
&= \sum_{t=p+1}^{T} \left\{ E_t \left[\tau_{\alpha_k} \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) \right] + Var_t \left[\tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) \right] \right\},
\end{aligned}$$

$$\begin{aligned}
\ddot{\ell}_{\alpha_m\alpha_n} &= \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \alpha_m \partial \alpha_n} \\
&= \sum_{t=p+1}^{T} \left\{ \frac{\frac{\partial^2}{\partial \alpha_m \partial \alpha_n} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \right. \\
&\quad \left. - \frac{\frac{\partial}{\partial \alpha_m} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \frac{\frac{\partial}{\partial \alpha_n} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \right\} \\
&= \sum_{t=p+1}^{T} \left\{ E_t \left[\tau \left(\alpha_m \diamond X_{t-m}; X_{t-m}, \alpha_m \right) \tau \left(\alpha_n \diamond X_{t-n}; X_{t-n}, \alpha_n \right) \right] \right. \\
&\quad \left. - E_t \left[\tau \left(\alpha_m \diamond X_{t-m}; X_{t-m}, \alpha_m \right) \right] E_t \left[\tau \left(\alpha_n \diamond X_{t-n}; X_{t-n}, \alpha_n \right) \right] \right\} \\
&= \sum_{t=p+1}^{T} Cov_t \left[\tau \left(\alpha_m \diamond X_{t-m}; X_{t-m}, \alpha_m \right), \tau \left(\alpha_n \diamond X_{t-n}; X_{t-n}, \alpha_n \right) \right].
\end{aligned}$$

$$\begin{aligned}
\ddot{\ell}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} &= \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} \\
&= \sum_{t=p+1}^{T} \left\{ \frac{\frac{\partial^2}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \right. \\
&\quad \left. - \frac{\frac{\partial}{\partial \boldsymbol{\lambda}} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \frac{\frac{\partial}{\partial \boldsymbol{\lambda}'} h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})}{h(X_t|\mathbf{X}_{-p};\boldsymbol{\theta})} \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=p+1}^{T} \left\{ E_t \left[\boldsymbol{\gamma}_{\boldsymbol{\lambda}} \left(\varepsilon_t; \boldsymbol{\lambda} \right) + \boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right)' \right] - E_t \left[\boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \right] E_t \left[\boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right)' \right] \right\} \\
&= \sum_{t=p+1}^{T} \left\{ E_t \left[\boldsymbol{\gamma}_{\boldsymbol{\lambda}} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \right] + Var_t \left[\boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \right] \right\},
\end{aligned}
$$

$$
\begin{aligned}
\ddot{\ell}_{\alpha_k \boldsymbol{\lambda}} &= \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \alpha_k \partial \boldsymbol{\lambda}} \\
&= \sum_{t=p+1}^{T} \left\{ \frac{\frac{\partial^2}{\partial \alpha_k \partial \boldsymbol{\lambda}} h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})}{h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})} \right. \\
&\quad \left. - \frac{\frac{\partial}{\partial \alpha_k} h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})}{h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})} \frac{\frac{\partial}{\partial \boldsymbol{\lambda}} h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})}{h(X_t | \mathbf{X}_{-p}; \boldsymbol{\theta})} \right\} \\
&= \sum_{t=p+1}^{T} \left\{ E_t \left[\tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) \boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \right] \right. \\
&\quad \left. - E_t \left[\tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right) \right] E_t \left[\boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \right] \right\} \\
&= \sum_{t=p+1}^{T} Cov_t \left[\tau \left(\alpha_k \diamond X_{t-k}; X_{t-k}, \alpha_k \right), \boldsymbol{\gamma} \left(\varepsilon_t; \boldsymbol{\lambda} \right) \right].
\end{aligned}
$$

■

Proof (of Theorem 3.3) We sketch the details as the proof follows from a standard Taylor series and remainder argument i.e.

$$
\dot{\ell}_{\boldsymbol{\theta}} \left(\hat{\boldsymbol{\theta}} \right) = \mathbf{0} = \dot{\ell}_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}_0 \right) - \ddot{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}} \left(\boldsymbol{\theta}_0 \right) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) + R(\boldsymbol{\theta}^*)
$$

where $\boldsymbol{\theta}_0$ is the true value of the parameter and R is a remainder term evaluated at $\boldsymbol{\theta}^*$, a convex linear combination of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. From DL p130 we know that X_t following an $INAR(p)$-P process is stationary and ergodic. Since the conditional expectations $E_t[\cdot]$ of the score and information matrix can be written as explicit functions of the process X_t (see Appendix B) it follows that these processes too are stationary and ergodic. The score function is also a martingale (and the sum of a martingale difference sequence) as it is the derivative of a scalar likelihood. From Azzalini (1983) p382 it follows that the MLE's are consistent. Since the elements of the information matrix are stationary and ergodic it follows from

consistency that the scaled remainder term in the Taylor series expansion of the score function is asymptotically negligible using a uniform law of large numbers. Thus the properties of the score determine those of the MLE's. i.e. scaling and solving we get

$$T^{1/2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=\left(T^{-1}\ddot{\ell}_{\boldsymbol{\theta\theta}}\right)^{-1}T^{-1/2}\dot{\ell}_{\boldsymbol{\theta}}+o_p\left(1\right).$$

Take an arbitrary linear combination of the score $l'\dot{\ell}_{\boldsymbol{\theta}}$. Since the score is the sum of a stationary, ergodic martingale difference sequence (with finite variance), $T^{-1/2}l'\dot{\ell}_{\boldsymbol{\theta}}$ automatically satisfies a univariate central limit theorem (see Davidson (1994) p385) and this linear combination is asymptotically normal. Using the Cramer-Wold device the proof is completed by showing that the score has finite variance and that the information matrix is non-singular. The mapping theorem then delivers the asymptotic distribution of $T^{1/2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)$. These steps are shown in detail by Freeland (1998) for $p=1$ and Bu (2006) for $p\geq 1$. ■

Appendix B: Time t Conditional Expectations for $INAR(p)$-P Process

For the $INAR(p)$-P process, the time t conditional expectations are functions of the transition probability. For example, it follows from Theorem 2.1 and the result in (14) that

$$\begin{aligned}
& E_t\left[\alpha_1\circ X_{t-1}\right] \\
= & \frac{\sum_{i_1=0}^{\min(X_{t-1},X_t)} i_1\binom{X_{t-1}}{i_1}\alpha_1^{i_1}(1-\alpha_1)^{X_{t-1}-i_1}P(X_t-i_1|X_{t-2},\ldots,X_{t-p})}{P(X_t|X_{t-1},\ldots,X_{t-p})} \\
= & \frac{\sum_{i_1=0}^{\min(X_{t-1},X_t)} X_{t-1}\binom{X_{t-1}-1}{i_1-1}\alpha_1^{i_1}(1-\alpha_1)^{X_{t-1}-i_1}P(X_t-i_1|X_{t-2},\ldots,X_{t-p})}{P(X_t|X_{t-1},\ldots,X_{t-p})} \\
= & \alpha_1 X_{t-1}\sum_{i_1=0}^{\min(X_{t-1},X_t)}\left\{\binom{X_{t-1}-1}{i_1-1}\alpha_1^{i_1-1}(1-\alpha_1)^{(X_{t-1}-1)-(i_1-1)}\right. \\
& \left.\times P(X_t-i_1|X_{t-2},\ldots,X_{t-p})\right\}\frac{1}{P(X_t|X_{t-1},\ldots,X_{t-p})}
\end{aligned}$$

$$
\begin{aligned}
&= \alpha_1 X_{t-1} \sum_{i_1=0}^{\min(X_{t-1}-1, X_t-1)} \left\{ \binom{X_{t-1}-1}{i_1} \alpha_1^{i_1} (1-\alpha_1)^{(X_{t-1}-1)-i_1} \right. \\
&\quad \left. \times P(X_t - 1 - i_1 | X_{t-2}, \ldots, X_{t-p}) \right\} \frac{1}{P(X_t | X_{t-1}, \ldots, X_{t-p})} \\
&= \frac{\alpha_1 X_{t-1} P(X_t - 1 | X_{t-1} - 1, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})}.
\end{aligned}
$$

By applying the same reasoning to (9), the following results hold.

$$
E_t \left[\alpha_k \circ X_{t-k} \right] = \frac{\alpha_k X_{t-k} P(X_t - 1 | X_{t-1}, \ldots, X_{t-k} - 1, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})},
$$

$$
E_t[\varepsilon_t] = \frac{\lambda P(X_t - 1 | X_{t-1}, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})},
$$

$$
\begin{aligned}
&\left\{ E_t \left[(\alpha_k \circ X_{t-k})^2 \right] - E_t \left[\alpha_k \circ X_{t-k} \right] \right\} \\
&= \frac{\alpha_k^2 X_{t-k} (X_{t-k} - 1) P(X_t - 2 | X_{t-1}, \ldots, X_{t-k} - 2, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})},
\end{aligned}
$$

$$
\left\{ E_t[\varepsilon_t^2] - E_t[\varepsilon_t] \right\} = \frac{\lambda^2 P(X_t - 2 | X_{t-1}, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})},
$$

$$
\begin{aligned}
&\left\{ E_t \left[(\alpha_m \circ X_{t-m}) (\alpha_n \circ X_{t-n}) \right] \right\} \\
&= \frac{\alpha_m \alpha_n X_{t-m} X_{t-n} P(X_t - 2 | X_{t-1}, \ldots, X_{t-m} - 1, \ldots, X_{t-n} - 1, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})},
\end{aligned}
$$

$$
\left\{ E_t[(\alpha_k \circ X_{t-k}) \varepsilon_t] \right\} = \frac{\alpha_k \lambda X_{t-k} P(X_t - 2 | X_{t-1}, \ldots, X_{t-k} - 1, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})}.
$$

See Bu (2006) for more details.